U.S. Army Research Institute
for the Behavioral and Social Sciences

Research Report 1549

Development of a Peer Comparison Procedure for the U.S. Army Aviation Officer Advanced Course

D. Michael McAnulty
Anacapa Sciences, Inc.

February 1990

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

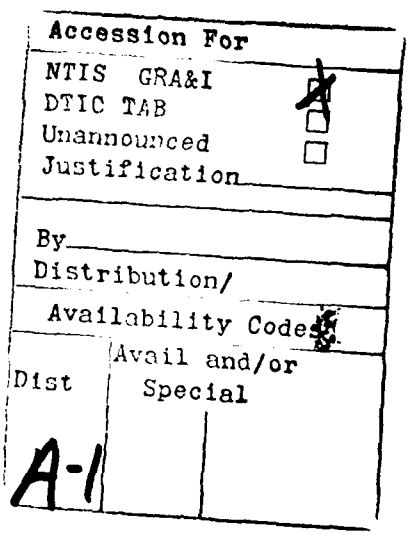
JON W. BLADES
COL, IN
Commanding

Research accomplished under contract
for the Department of the Army

Anacapa Sciences, Inc

Technical review by

N. Joan Blackwell
Charles A. Gainer
Gabriel P. Intano
Ronald J. Lofaro



NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POX, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS --
2a. SECURITY CLASSIFICATION AUTHORITY --	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE --	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) ASI690-320-89	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Research Report 1549
6a. NAME OF PERFORMING ORGANIZATION Anacapa Sciences, Inc. 6b. OFFICE SYMBOL (If applicable) --	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute Aviation Research and Development Activity
6c. ADDRESS (City, State, and ZIP Code) P.O. Box 489 Fort Rucker, Alabama 36362-5000	7b. ADDRESS (City, State, and ZIP Code) ATTN: PERI-IR Fort Rucker, Alabama 36362-5354
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences 8b. OFFICE SYMBOL (If applicable) PERI	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903-87-C-0523
8c. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO.: 63007A PROJECT NO.: 792 TASK NO.: 2204 WORK UNIT ACCESSION NO.: C6

11. TITLE (Include Security Classification)
Development of a Peer Comparison Procedure for the U.S. Army Aviation Officer Advanced Course

12. PERSONAL AUTHOR(S)
McAnulty, D. Michael (Anacapa Sciences)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM 86/10 TO 89/01	1990, February	

16. SUPPLEMENTARY NOTATION All research on this project was technically monitored by Mr. Charles A. Gainer, Chief, U.S. Army Research Institute Aviation Research and Development Activity (ARIARDA), Fort Rucker, Alabama.

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	08	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Peer nomination; Peer ranking; Paired comparisons; Army aviation officer; Military qualities; [illegible] selection

19. ABSTRACT (Continue on reverse if necessary and identify by block number)
In response to a request from the School Secretary, U.S. Army Aviation Center, Fort Rucker, Alabama, a peer comparison (PC) procedure was developed to select Aviation Officer Advanced Course (AVNOAC) honor graduates on the basis of a "whole person" concept. Under this concept, students would be evaluated on and honored for both their academic performance and for other attributes important in the development of an Army aviation officer. The PC procedure is a combination of the peer nomination and peer ranking methods and the psychophysical scaling technique of paired comparisons. Section members in each AVNOAC class nominate and rank order five of their peers on their potential as aviation officers. They then compare each pair of nominees on each of five military qualities identified by senior aviation officers as important and likely to be observed in the course. The PC procedure was administered twice in each of two AVNOAC classes (N = 90 and 103). The results indicate that the procedure is easy to use, has high internal consistency and temporal stability, and produces a consensus on which students have the highest potential.

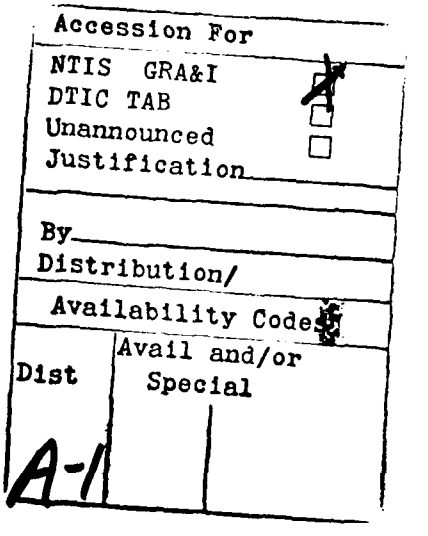
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Charles A. Gainer, COTR	22b. TELEPHONE (Include Area Code) (205) 255-4404 22c. OFFICE SYMBOL PERI-IR

DD Form 1473, JUN 86 *Previous editions are obsolete.* SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

Research Report 1549

Development of a Peer Comparison Procedure for the U.S. Army Aviation Officer Advanced Course

D. Michael McAnulty
Anacapa Sciences, Inc.



Aviation R&D Activity at Fort Rucker, Alabama
Charles A. Gainer, Chief

Training Research Laboratory
Jack H. Hiller, Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

February 1990

Army Project Number
2Q263007A792

Manpower and Personnel

FOREWORD

The Army Research Institute Aviation Research and Development Activity (ARIARDA) at Fort Rucker, Alabama, is responsible for providing timely research and development support in aircrew training for the U.S. Army Aviation Center (USAAVNC). Research and development activities are conducted in-house and augmented by contract support as required. This technical report documents contract work performed by ARIARDA in support of the School Secretary's Office at the USAAVNC. The research was initiated as a technical advisory service in response to a request from the School Secretary in February, 1985, and was conducted as part of the aviator selection, assignment, and retention program at ARIARDA.

The successful development of Army aviation officers requires high standards of performance in training and in the units. There is a need to identify junior officers with high potential to ensure that they receive the training and experience needed to qualify them for senior command positions. To meet these requirements, the School Secretary requested that ARIARDA develop an evaluation procedure that would (a) motivate students to maximize their military and academic efforts during the Aviation Officers Advanced Course, and (b) identify students who have high potential as Army aviation officers early in their careers. This report describes the peer comparison (PC) procedure that was developed to meet the School Secretary's request and the results of two experimental administrations of the procedure in the Advanced Course.

This report meets two objectives. First, it provides all of the materials and information needed to evaluate and implement the PC procedure in the military training courses at the USAAVNC or at other Army installations. Second, it provides summary information on the PC procedure to behavioral scientists working on similar applied research issues in other governmental, industrial, or university organizations. The results of this research have been briefed to the Director and representatives of the School Secretary's Office and the Directorate of Aviation Proponency at the USAAVNC. The implementation of the PC procedure is being considered for several courses at the USAAVNC.

EDGAR M. JOHNSON
Technical Director

ACKNOWLEDGMENTS

The research reported here was performed at Fort Rucker, Alabama, by personnel from the U.S. Army Aviation Center School Secretary's Office, the Army Research Institute Aviation Research and Development Activity (ARIARDA), and Anacapa Sciences, Inc. The author would like to thank the following individuals.

COL Norman Ferguson, now retired, was the School Secretary who initiated this research project. Three Assistant School Secretaries, MAJ Charles Hersant, CPT James Collins, and CPT Langford Fowler coordinated the experimental administrations and assisted in collecting and processing the peer comparison data.

Mr. Charles A. Gainer, ARIARDA chief, served as the Contracting Officer's Technical Representative throughout the project and provided guidance on many aspects of the research. MAJ Lynn Hansen of ARIARDA helped coordinate the experimental administrations.

Dr. Kenneth Cross, Anacapa Sciences, Inc., participated in the initial formulation of the peer comparison approach. Mr. Theodore Aldrich and Dr. John Ruffner conducted a critical review of an earlier draft of this report. Ms. Carol Lynde and Ms. Jenny Grimes assisted in the data processing and Ms. Nadine McCollim produced the original data collection materials and this project report.

DEVELOPMENT OF A PEER COMPARISON PROCEDURE FOR THE U.S. ARMY AVIATION OFFICER ADVANCED COURSE

EXECUTIVE SUMMARY

Requirement:

This research was initiated in response to a request from the School Secretary of the U.S. Army Aviation Center, Fort Rucker, Alabama, to develop a new method of selecting aviation officer course graduates for honors on the basis of a "whole person" concept. Under this concept, the outstanding students in a course would be evaluated on and honored for both their academic performance and for other attributes that are important in the development of an Army aviation officer. The purpose of the new method was (a) to motivate students to maximize their military and academic efforts during the course, and (b) to identify students who have high potential as Army aviation officers.

Procedures:

Senior Army aviation officers were surveyed to identify the five military qualities that were most important to the performance of captains and senior aviation officers and most likely to be demonstrated during the Aviation Officer Advanced Course (AVNOAC). The AVNOAC is a 5-month officer training course for captains and promotable first lieutenants. The five qualities selected for evaluation were (a) leadership, (b) responsibility, (c) communication, (d) appearance, and (e) cooperation. A peer comparison (PC) procedure was developed that is a combination of the peer nomination and peer ranking methods and the psychophysical scaling technique of paired comparisons. On the PC form, section members in each class were asked to nominate and rank order five of their peers on their potential as aviation officers. The section members were then asked to compare each pair of nominees on each of five military qualities. The PC procedure was administered twice in each of two AVNOAC classes (N = 90 and 103). PC data were collected during the course (after 4 months and 2 months, respectively, in the two classes) and again at the end of the course. Student critiques and faculty advisor ratings were also collected at the end of the course.

Findings:

The results indicate that the PC procedure is easy to use and produces highly reliable results, both in terms of the

internal consistency of the components of the PC procedure and the stability of the evaluations over 1- and 3-month periods. The results from each class section exhibited a consensus among the members about which peers have the highest potential as Army aviation officers. Anecdotal reports from the students indicated that the peer evaluations caused many of the class members to improve their military decorum during the course. However, the class members generally had a negative reaction to the PC procedure. Approximately 70 percent of the students were opposed to the implementation of the procedure in the AVNOAC.

Utilization:

Despite the negative peer reactions, the research results support the implementation of the PC procedure in the AVNOAC course. Two methods of combining academic grades and PC scores (a multiple gate approach and a weighted sum approach) are discussed using different cutoffs and weights for the PC scores. Technical recommendations are made for implementing the PC procedure in the AVNOAC for a 1-year trial period.

DEVELOPMENT OF A PEER COMPARISON PROCEDURE FOR THE U.S. ARMY AVIATION OFFICER ADVANCED COURSE

CONTENTS

CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

GLOSSARY OF ACRONYMS AND ABBREVIATIONS

APPR	-	Appearance
ARIARDA	-	Army Research Institute Aviation Research and Development Activity
AVG	-	Average
AVNOAC	-	Aviation Officer Advanced Course
COMM	-	Communication
COOP	-	Cooperation
FAR	-	Faculty Advisor Rating
LDRS	-	Leadership
NOM	-	Number of Nominations
PC	-	Peer Comparison
RESP	-	Responsibility
USAAVNC	-	U.S. Army Aviation Center

DEVELOPMENT OF A PEER COMPARISON PROCEDURE FOR THE U.S. ARMY AVIATION OFFICER ADVANCED COURSE

INTRODUCTION

This research was initiated in response to a request from the School Secretary of the U.S. Army Aviation Center, Fort Rucker, Alabama. The School Secretary requested support in developing a new method to select aviation officer course graduates for class honors on the basis of a "whole person" concept. At that time, course graduates were awarded honors solely on the basis of their academic performance. Under the whole person concept, the outstanding students in a course would be evaluated on and honored not only for academic achievements but also for other attributes that are important in the development of Army aviation officers.

As an initial test of the whole person concept, the School Secretary wanted to augment the academic grade criterion used to select graduates for honors in the Aviation Officer Advanced Course (AVNOAC). The primary purposes of the augmented program are:

- to motivate students to maximize their military as well as their academic efforts during the course, and
- to identify students who have high potential as Army aviation officers at an early stage of their careers.

AVNOAC Description

The AVNOAC is a 5-month officer training course for recently promoted captains and promotable first lieutenants. The course is designed to prepare the students for successful performance in company command positions. Each class has approximately 100 students divided into two sections. When this research was conducted, the majority of the course was taught with classroom lectures and demonstrations by subject matter experts.

At that time, student performance was evaluated by 25 standardized academic examinations. The distinguished graduate, honor graduates, and Commandant's list graduates in each class were determined solely on the basis of the academic average. These awards were entered into each student's permanent military records, which are used to make assignment and promotion decisions.

Evaluation Alternatives

There were three possible sources of nonacademic evaluations in the AVNOAC: course instructors, faculty advisors, and class peers. As discussed above, the course instructors usually taught only a few class sessions in their areas of expertise and usually used a lecture approach. This procedure limited the instructors' opportunities to observe and evaluate the students on other than academic performance. The limited interaction between the course cadre and the students precluded the use of instructor evaluations to augment the academic criterion.

Faculty advisors. The second potential source of student evaluations was the AVNOAC faculty advisors. However, the information the advisors could provide was limited by two factors. First, each advisor was usually assigned only five to eight students and had very little exposure to the entire class. Second, the advisors usually did not meet with their advisees on a regular basis. Instead, the advisors were available for student-initiated or instructor-referred counseling. As a result, it was assumed that faculty advisor evaluations would be based primarily on academic performance and possibly on negative information (e.g., lack of progress, misconduct, personal problems) that precipitated or was discussed during counseling sessions.

Because of the lack of opportunity for faculty advisors to observe all the students regularly during the course, faculty advisor ratings were not considered appropriate input for the honor graduate algorithm. The opportunity to observe the ratees is a basic requirement for producing valid evaluations (Kane & Lawler, 1978). However, it was proposed that faculty advisor ratings be collected to test the assumption that the evaluations would be highly correlated with academic performance.

Peer evaluations. The third source of evaluative information was the class peers. Kane and Lawler (1978) suggested that peer evaluations are likely to be effective when three conditions are met:

- the peers have the opportunity to observe salient aspects of each other's behavior,
- the group members are capable of perceiving and interpreting the salient aspects of behavior, and
- the need for the evaluation is apparent to the group members.

Within each AVNOAC section, the students interact during the class sessions, participate in class demonstrations and

exercises together, and presumably engage in social activities together outside of the course environment. This level of interaction provides sufficient opportunities for the group members to observe the behavior of their peers in several different contexts. As military officers, the class members are accustomed to being evaluated by their seniors and are trained to observe and evaluate their subordinates, thus satisfying the second condition. Finally, as officers in a hierarchical organization performing a complex mission, it is assumed that the students understand the need for identifying and providing appropriate training for individuals who have the highest potential for successful performance in senior positions.

As a result of these considerations, the School Secretary directed that this research should concentrate on the development and evaluation of a peer assessment procedure for use in the AVNOAC. The research that was conducted to meet these objectives is described in the following major sections of this report:

- review of the peer assessment literature,
- methods of peer assessment,
- development of evaluation forms,
- preliminary administration,
- second administration,
- utilization options, and
- summary and recommendations.

REVIEW OF THE PEER ASSESSMENT LITERATURE

Peer assessment is the process of having group members make judgments about specified traits, behaviors, or attributes of other members of the group (cf. Kane & Lawler, 1978). In varying degrees, past reviews of peer assessment research (e.g., DeNisi & Mitchell, 1978; Downey & Duffy, 1978; Kane & Lawler, 1978; and Lewin & Zwany, 1976) and reviews of performance appraisal or prediction that have included peer assessments (e.g., Griffin & Mosko, 1977; Harris & Schaubroeck, 1988; and Korman, 1968) reached positive conclusions about the potential utility of peer assessments. For example, Downey and Duffy concluded that "peer evaluations are a powerful tool in discriminating complex human behavior" (p. 19). In investigating naval aviator attrition, Griffin and Mosko concluded that "Of all measures studied...peer ratings and instructor ratings have been shown to be consistently powerful predictors of success and failure" (p. 18). Korman concluded that the "peer rating paradigm is the most consistently effective predictor of military officer behavior" (p. 313).

These reviews agreed that the major peer assessment methods consistently exhibit acceptable levels of reliability and validity when employed in industrial and military organizations. The research indicated that valid peer assessments are formed quickly and are not severely affected by many interpersonal and situational variables. For example, Hollander (1957) found that peer assessments stabilized within three weeks and that there was no difference in reliability between assessments made for research purposes only and for administrative purposes. Love (1981) found that friendship bias did not affect the reliability and validity of peer assessments; Kane and Lawler (1978) suggested that the relationship between friendship and peer assessments reflects a tendency to choose friends who are very capable rather than a tendency for friendship to influence judgments. Finally, Harris and Schaubroeck (1988) found that the relationship between peer and supervisor evaluations are not moderated by rating format (dimensional versus global), scale type (trait versus behavioral), or job type (professional/ managerial versus blue-collar/service).

Research With Army Officers

A majority of the peer assessment research has been conducted in military settings (Kane & Lawler, 1978) and in training situations (Downey & Duffy, 1978). Several recent peer assessment evaluations have been conducted that are

highly germane to the present research context. For example, Downey, Medland, and Yates (1976) found that peer assessments by senior U.S. Army officers (colonels) within the same career field exhibited high split-half reliability coefficients and significantly predicted promotion to brigadier general. At a lower officer level, Gilbert and Downey (1978) found that peer assessments obtained during Army Ranger training were significant predictors of officer ratings on ten performance dimensions collected 3 years later.

In an Army aviation context, Wahlberg, Boyles, and Boyd (1971) found that peer assessments during the Aviation Warrant Officer Candidate Military Development Course were significant predictors of success in primary flight training. Finally, Eastman and his associates (Eastman & Leger, 1978; Eastman & McMullen, 1976) demonstrated the reliability and validity of peer assessments in selecting pilots for the AH-1 helicopter transition course.

Problems With Peer Assessment

Despite the consistently positive psychometric evaluations of peer assessments, there are several problems associated with their use. These problems are manifested in the limited operational use of the peer assessment methods (Kane & Lawler, 1978; McEvoy & Buller, 1987). Several of the general problems are discussed in this section; problems that are specific to the different peer assessment techniques are discussed in the next section.

As previously mentioned, a moderate relationship has frequently been found between friendship and peer assessments (e.g., Hollander, 1954; 1956; Hollander & Webb, 1955; Waters & Waters, 1970), but this relationship is not thought to invalidate the peer judgments (Kane & Lawler, 1978; Love, 1981). Cox and Krumboltz (1958) and deJung and Kaplan (1962) found that peer ratings were highly racially biased. More recently, however, Schmidt and Johnson (1973) found no evidence of a racial bias effect with peer assessments. Nonetheless, the personal bias or "popularity contest" issue is the most consistent argument against the operational use of peer assessments (Downey & Duffy, 1978). Although the recent research indicates that these interpersonal and demographic factors do not significantly affect the validity of peer assessments, the perceived effect of these factors probably influences the acceptability of the technique to potential users.

McEvoy and Buller (1987) suggest that low user acceptance may be a major reason that peer assessments are not more widely used. Very few investigations have addressed the issue of user acceptance, and some of these only indirectly (e.g., Downey, Medland, & Yates, 1976). Most of the research that has been conducted found generally negative reactions from the users (e.g., Cederblom & Lounsbury, 1980; Downey, et al.; Love, 1981) or a neutral reaction (Mayfield, 1970). Only two studies have reported positive user reactions (McEvoy & Buller; Roadman, 1964). In both investigations, the most positive reactions were associated with peer appraisals that were used for developmental rather than administrative purposes. Love did not find any difference among the peer assessment methods in user acceptability.

There is an obvious need to evaluate the reaction of group members toward any peer assessment procedure before implementing it for operational use. Depending upon the peer assessment method used, group members who are highly opposed to its use can easily "game" the appraisal and subvert its effectiveness. Therefore, the present research will attempt to evaluate in detail the AVNOAC users' reactions to the peer assessments.

There are also several technical problems (e.g., difficulty in design, administration, and scoring) that are specific to each peer assessment method. Each of these problems and between-method differences in reliability and validity are discussed in the descriptions of the major peer assessment techniques.

METHODS OF PEER ASSESSMENT

Several peer assessment procedures, combinations of procedures, and situationally specific assessment dimensions and administrative conditions have been described in the literature. However, most reviewers group peer assessments into three primary methods: peer ratings, peer nominations, and peer rankings (e.g., Downey & Duffy, 1978; Kane & Lawler, 1978; Love, 1981). Each of the primary methods, and a fourth method (paired comparisons) that has been used in only a few studies, are described in the following sections.

Peer Rating

Peer rating is similar to the widely used supervisor rating procedure, except for the number and organizational position of the individuals providing the ratings. In this method, group members rate their peers on a specified set of personal characteristics or behaviors using a variety of rating scales. The major advantages of the rating procedure are that (a) data are obtained about all members of the group, (b) scoring and combining ratings across raters or dimensions is relatively straightforward, and (c) the resulting data are usually assumed to be interval level measures.

There are several disadvantages to the rating procedure. First, the psychometrically superior measurement scales, such as behaviorally anchored rating scales, are expensive and time consuming to develop. Second, peer ratings are susceptible to nearly all forms of rater bias (see Bernardin & Beatty, 1984; Saal, Downey, & Lahey, 1980). The most common and damaging biases in terms of diminishing the utility of the data are the tendencies of raters to be lenient in their ratings (leniency bias), to rate almost all members the same (uniformity bias), and to not discriminate among the rating dimensions (halo effect). As a result, peer ratings have exhibited the lowest levels of reliability and validity of any of the peer assessment procedures.

In the AVNOAC evaluation context data are only required about the most outstanding students in each class, and the potential for high reliability and validity is important for predicting such a long-term criterion as senior officer performance. The assumption that ratings produce interval level data has also been questioned (McAnulty & Jones, 1984). Given these considerations, the investment required to produce highly effective rating scales was judged not to be warranted. Although widely used, peer ratings were not

considered further as a component of the AVNOAC honors algorithm.

Peer Nomination

Peer nomination has been the most widely used and researched of the peer assessment methods (Kane & Lawler, 1978). In this method, each member of the group is asked to name a specified number of peers who are the highest or lowest (or both) on one or more characteristics or behaviors. The research evidence indicates that the nomination method results in very high levels of reliability and validity, although the results are better for positive (i.e., high only) than negative nominations. The coefficients may be artificially high, however, because the procedure only involves the extreme high and low members of the group. The nomination method is also easy for the group members to use.

In contrast, the method is complicated in design and scoring. For example, differences in group size must be considered in determining the number of nominations to request. Group size, the appropriateness of mathematical operations with nominal data, and the treatment of positive and negative nominations complicate the scoring process. Furthermore, the method provides no information about group members who are not in the high or low group and there is no discrimination between the nominations made by each rater. Finally, peer nomination is the method that is most susceptible to friendship or subgroup bias.

The positive aspects of the peer nomination method are suitable for the purposes of this research. The major advantage of the method is its high levels of reliability and validity, which are extremely important for predicting long-term performance. The method is also easy to use. Furthermore, some of the general disadvantages of the method do not affect the current situation: data are only required about the extreme high individuals in each class and the class sections are approximately the same size. However, it is desirable to obtain more precise data that discriminate among the nominees and are less susceptible to bias effects.

Peer Ranking

Peer ranking is the least researched of the three major peer assessment methods. In this method, each group member rank orders the other members from best to worst on one or more characteristics or behaviors. If no ties are allowed,

ranking is the most discriminating assessment method because each peer receives a different assessment score. That is, each member of the group is evaluated and no two members are given the same rank. The method is relatively simple to design and administer, and the limited available evidence indicates that peer ranking has satisfactory levels of reliability and validity (Kane & Lawler, 1978).

Three principal problems are associated with the peer ranking method. First, it is difficult to use with large or homogeneous groups. Second, the ordinal data produced by peer rankings make it difficult to derive mathematically sound composite scores for the group members. Third, the data provide no information about the distance on the evaluation scale between members having consecutive ranks.

The rank ordering of peer nominees is one of the more common combinations of peer assessment methods (Kane & Lawler, 1978), and it is suitable for the AVNOAC situation. This combination identifies the extreme high members of the group and provides a more precise level of discrimination by assigning a different score to each nominee. However, the combination still does not indicate the interval between similarly ranked nominees.

Paired Comparisons

The paired comparison method is a basic psychophysical scaling technique that is "often regarded as the most appropriate way of securing value judgments" (Engen, 1971, p. 51). In this method, each pair of stimuli is presented to each subject who must judge which stimulus is higher in value. The judgments can be used to derive an equal interval scale. The primary limitation of the method is that $N(N-1)/2$ pairs of stimuli must be presented; as the number of stimuli increase beyond ten or so, the number of comparisons becomes cumbersome and tedious.

Three studies were found that used paired comparisons to collect peer assessments. Lawshe, Kephart, and McCormick (1949) concluded that pair comparisons produced ratings that reflect the subjectively perceived distance between the persons evaluated. Bolton (1971) found that the reliability of paired comparisons was stable over a 2-week interval for all the squads participating in an Army Reserve Officer Training Corp summer camp. Most recently, Siegel (1982) used the paired comparison method to collect peer and supervisor evaluations for 20 savings and loan branch managers. The evaluations were part of the information considered in making

promotion decisions. Siegel found that the paired comparison appraisals exhibited high levels of interjudge agreement between and within the two evaluator groups.

Although not widely used for peer appraisals, the paired comparison procedure may potentially redress some of the problems associated with the nomination and ranking methods (Siegel, 1982). Specifically, the paired comparison procedure may provide better discrimination among nominees and produce data that may be more appropriately combined across raters. Having each rater compare only his or her own nominees limits the number of required comparisons to a manageable number.

Peer Comparison Concept

The characteristics of the AVNOAC evaluation context and the review of the peer assessment methods led to the development of the peer comparison (PC) concept. The PC concept is a combination of the peer nomination, peer ranking, and paired comparison methods. Under this concept, AVNOAC section members will first nominate five of their peers who are judged to have the highest potential as Army aviation officers. Each section member will then rank order the nominees in terms of their potential. Finally, each section member will make paired comparisons between the nominees on five military qualities that are important in the development of aviation officers.

DEVELOPMENT OF EVALUATION FORMS

Three data collection forms were developed for this research: the PC form, the faculty advisor rating (FAR) form, and the student critique form. The general concept of the PC procedure was derived from the literature on peer assessment methods. Before the the PC form was developed specifically for the AVNOAC situation, however, it was necessary to determine the military quality dimensions that would be evaluated by the class peers.

Military Qualities Survey

Following a search of the literature (e.g., Burke, Kramer, & Butler, 1982; Puryear, 1971; Rogers, Lilley, Wellins, Fischl, & Burke, 1982) and a review of current Army student evaluation dimensions (e.g., officer evaluation reports), the definitions of 14 primary military qualities were compiled for consideration as evaluation dimensions by senior aviation officers. Several important military qualities were excluded from the survey because they are evaluated by academic scores (e.g., technical competence) or are unlikely to be demonstrated during the AVNOAC (e.g., development of subordinates). The 14 military qualities included in the survey and their definitions are listed in Table 1.

Sixteen senior Army aviation officers at Fort Rucker were asked to rate each of the 14 military qualities on the following four scales:

- importance to the performance of captains,
- importance to the performance of senior officers,
- probability of demonstration during the AVNOAC, and
- degree of overlap with the other qualities.

On each scale, the scores can range from 0 to 99. A score of 0 indicates not at all important, no probability of demonstration, or no overlap, respectively. A score of 99 indicates the quality was critically important, extremely probable, or identical to another quality, respectively. If a rating on the fourth scale was greater than 50, the respondent was asked to identify the qualities that overlapped.

On each scale, the respondents were asked to assign a different score to each military quality. In addition, the respondents were asked to reflect the relative value of each quality in its assigned scale score. For example, if one quality was judged to be twice as important as another, the respondents were told to rate the first quality twice as high

Table 1

Military Quality Definitions

Quality	Definition
Adaptability:	Performs effectively despite changes in personnel, resources, or circumstances; seeks self-improvement to meet changing conditions.
Analysis:	Identifies problems, secures relevant information, integrates data from different sources, and identifies possible problem causes.
Appearance:	Maintains a military appearance and bearing in dress, grooming, posture, carriage, and physical fitness that instills confidence and respect in others.
Communication:	Expresses ideas clearly and effectively both orally and in writing; utilizes good grammatical form and appropriate gestures to enhance accurate communication.
Cooperation:	Acts in concert with others to achieve mutual goals; subordinates personal objectives to the goals of the group or organization.
Decisiveness:	Makes decisions, renders judgments, or takes action in a timely manner without needlessly seeking further information.
Delegation:	Uses subordinates effectively; allocates decision making and other responsibilities to the appropriate subordinates.
Initiative:	Attempts to influence events to achieve goals rather than passively accepting events; originates action to achieve goals beyond those that are minimally acceptable.

Continued on the next page

Table 1. Military Quality Definitions (Continued)

Quality	Definition
Judgment:	Develops alternative courses of action and makes decisions based on logical assumptions that reflect factual information.
Leadership:	Utilizes appropriate interpersonal styles and methods in guiding individuals (subordinates, peers, supervisors) or groups toward task accomplishment.
Organization:	Establishes courses of action for self and/or others to accomplish specific goals; plans proper assignment of personnel and allocation of resources.
Responsibility:	Completes duties in a timely, reliable, and effective manner; seeks authority for additional actions required to maintain or improve performance and readiness; accepts accountability for obligations and actions.
Sensitivity:	Indicates a consideration and concern for the feelings and needs of others as well as the needs of the organization.
Supervision:	Establishes procedures to monitor and/or regulate ongoing processes, tasks, or activities; takes action to maintain high standards of performance in delegated assignments or projects.

as the second (e.g., 20 and 10, or 90 and 45, depending on the absolute values of the two qualities). Finally, the respondents were asked to indicate the five qualities they would recommend for the PC form.

Fifteen surveys were returned, but only 11 were complete and usable. Although the number of usable surveys is small, they collectively represented the opinions of a majority of the senior aviation officers at Fort Rucker. Seven of the qualities (adaptability, analysis, decisiveness, delegation, organization, sensitivity, and supervision) received moderate

ratings on the first three scales and were rarely selected as one of the five qualities recommended for the PC form. These seven qualities were not considered further as PC dimensions.

Three of the qualities (leadership, judgment, and responsibility) had consistently high ratings on the first three scales and were selected as PC dimensions. However, leadership and judgment were rated as having a substantial overlap. Therefore, these two qualities were combined to form a single dimension of leadership defined as:

> Utilizes appropriate interpersonal styles and methods in guiding individuals or groups toward task accomplishment; exercises good judgment in developing alternative courses of action and in making decisions.

Of the remaining four qualities, communication, appearance, and cooperation were selected as the final three PC dimensions. Each of these qualities was rated to be moderately high on the first three scales and to have minimal overlap with the other qualities. Although rated as highly important to the performance of captains and senior officers, initiative was not selected as a PC dimension because it was rated as unlikely to be demonstrated and observed in the AVNOAC.

PC Form Development

The PC form (see Figure 1) was developed from (a) the results of the military qualities survey, (b) a combination of the peer nomination and peer ranking techniques, and (c) the psychophysical method of paired comparisons (e.g., Engen, 1971, pp. 51-54). The form was designed to be administered separately in each class section. On the PC form, each section member is required to nominate and rank order five peers on the basis of their potential as Army aviation officers. The section member then makes paired comparisons of the nominees on each of the five military qualities that were selected from the military qualities survey.

Scoring procedures. PC scores are computed for each peer by first summing the rank score (five points for first rank, four points for second rank, ..., one point for fifth rank) from each nominating section member. The summed rank scores are then added to the number of favorable comparisons

AVNOAC PEER COMPARISON FORM (EXPERIMENTAL)

1. Nominate and rank order the five members of your section, excluding yourself, who have the highest potential as U.S. Army aviation officers (list by PC Nr):

   1 (highest potential) - _____ 2 - _____

   3 - _____ 4 - _____ 5 - _____

2. Compare each nominee with the other nominees on the military qualities indicated below. On each comparison, write the rank order number of the nominee who is higher on that quality.

	MILITARY QUALITIES				
Compare Nominees	LDRS	RESP	COMM	APPR	COOP
1 vs 2					
1 vs 3					
1 vs 4					
1 vs 5					
2 vs 3					
2 vs 4					
2 vs 5					
3 vs 4					
3 vs 5					
4 vs 5					

3. PC Nr: _____ CLASS: _____ DATE: _____

Figure 1. The PC data collection form.

the peer received on each military quality. Finally, the total is divided by the maximum possible score to enable direct comparisons between sections with unequal numbers of students.

The PC scores can range from 0.0 (no nominations) to 1.0 (ranked first by all section members and always favorably compared with the other nominees). If all the PC judgments were made randomly, each section member would receive a PC of approximately .05. Alternatively, if the top five students were equal in potential and divided all the PC points, each of the five students would receive a PC score of .20. Although there is no precedent for determining what a "significant" PC score is, a PC of .20 (i.e., four times the randomly expected score) or greater probably represents a consensus among the section members that the student has high potential as an Army aviation officer.

Faculty Advisor Rating Form

A faculty advisor rating (FAR) form was developed to obtain independent evaluations of the students' potential as Army aviation officers. The faculty ratings were designed for comparison with the peer and academic evaluations rather than as part of the honors criterion. On the FAR, the faculty advisors were asked to estimate the officer potential of their students by assigning them percentile ranks in an average group of 100 captains (see Figure 2).

Student Critique Forms

Finally, a student critique form (see Appendix A) was developed to ascertain student attitudes toward the peer comparison program. The students were asked to rate the fairness, utility, aversiveness, and difficulty of various aspects of the program. They were also asked to express their opinions about the implementation of the program and to offer recommendations for improving the program.

AVNOAC FACULTY ADVISOR RATINGS

In an average class of 100 Captains attending the AVNOAC, what rating would you give your advisees in terms of their potential as U.S. Army aviation officers? (check one rating for each advisee.)

	WELL ABOVE AVERAGE		ABOVE AVERAGE		AVERAGE	BELOW AVERAGE		WELL BELOW AVERAGE	
ADVISEE NAME	TOP 1%	TOP 5%	TOP 10%	TOP 25%	MIDDLE 50%	LOW 25%	LOW 10%	LOW 5%	LOW 1%

Please comment on any extreme ratings or unusual circumstances:

Thank you for your assistance. The ratings you provide will be treated as confidential and will be used for research purposes only.

Figure 2. The faculty advisor rating (FAR) form.

PRELIMINARY ADMINISTRATION

As a preliminary test of the PC procedure, data were collected on an experimental basis from the class in residence (AVNOAC 85-2) as soon as the three evaluation forms were developed. PC data were collected twice in each section to determine the stability of the evaluations over time. The first data collection was at the end of the fourth month of the course. The second data collection was 1 month later at the end of the course. The FAR and student critique data were collected only at the end of the course.

Method

All the data were collected separately for each section during regularly scheduled class periods. Before the first data collection, the School Secretary gave a brief introduction and the researcher explained the PC objectives and procedures and answered student questions. In particular, the students were advised that the PC data were being collected for research purposes only and would not be used for administrative purposes (i.e., selecting students for class honors). After the introductory comments, instructions, and questions, the five military quality definitions, the PC form, and the class roster were distributed to the students. All students completed the PC forms within 10 minutes.

During the second data collection, the purpose and procedures of the research were briefly reviewed before the students began working on the PC evaluations. All students completed the PC forms within 10 minutes. An additional 10 minutes was required for the students to complete the critique forms.

Immediately after the class graduated, the faculty advisors completed the FARs and the final academic averages (AVGs) were obtained from the School Secretary's office.

Results

Usable peer comparisons were collected from 38 students in Section 1 and from 40 students in Section 2 at the end of the fourth month of training. Three additional PC evaluations from the students in Section 1 were not retained for analysis because they were not completed correctly. Of the 47 students in Section 1, 23 received more than one nomination; of the 43 students in Section 2, 22 received more than one nomination.

The second set of PC ratings and the student critiques were collected from 33 students in Section 1 and from 28 students in Section 2 at the end of the course, approximately 1 month later. During the second data collection, 25 students in Section 1 and 22 students in Section 2 received more than one nomination. Table 2 presents the frequency distributions of PC scores in each section for each data collection.

All the students received a FAR rating from their faculty advisors. The faculty advisor ratings tended to be very lenient, although there were a few very low scores. The median percentile was 75 in both sections, with a range of 25-99 in Section 1 and 10-95 in Section 2. Valid ratings would be expected to have a median percentile of 50. The mean and standard deviation of the AVGs for the 90 graduates reflected high overall performance and limited discrimination between students. The mean AVG was 92.9 (*SD* = 2.98) in Section 1 and 92.5 (*SD* = 3.11) in Section 2.

Table 2

Frequency Distribution of Peer Comparison (PC) Scores in Sections 1 and 2 of AVNOAC 85-2

PC Range	Section 1			Section 2		
	PC1	PC2	PC+	PC1	PC2	PC+
.000 - .049	32	35	34	28	27	28
.050 - .099	9	3	6	6	6	7
.100 - .149	0	3	1	4	5	2
.150 - .199	2	3	2	3	1	3
.200 - .249	1	1	1	0	2	1
.250 - .299	1	0	1	0	1	0
.300 - .349	0	0	0	2	0	1
.350 - .399	1	0	1	0	0	1
.400 - .449	0	1	0	0	1	0
.450 - .499	1	1	1	0	0	0

Note. There were 47 students in Section 1 and 43 students in Section 2. PC1 = first data collection; PC2 = second data collection; PC+ = combined PC scores.

The scores for the first and second data collections were highly correlated (Section 1 = .96 and Section 2 = .86), indicating the stability of the appraisals over time. Because of the high correlations, the ratings from the two data collections were combined into a single PC score for each peer. The combined PC scores ranged from .00 to .48 in Section 1 and from .00 to .36 in Section 2 (see Table 2). Four peers in Section 1 and three peers in Section 2 received PC scores greater than .20. The majority of the PC scores in both sections were between .00 and .05 (i.e., less than the random probability). The scores indicate a high consensus among the members of the class in identifying peers with the highest potential as aviation officers.

External correlations. The combined PC scores were then correlated with the FARs and AVGs. For Sections 1 and 2, respectively, the PC correlations were .45 and .33 with the FAR, and .55 and .30 with the AVG. These correlations are sufficiently high to show an expected relationship between observations of the same individuals. At the same time, the correlations are sufficiently low to indicate that the PC score was measuring a unique perspective of the class members. The correlations between the FAR and AVG were .76 and .59 in Sections 1 and 2, respectively. This result probably indicates that the faculty advisors were depending upon the academic average as a primary source of information in making their ratings.

PC critique. Finally, the responses to the PC critique were tabulated. The overall reaction of the class members to the PC program was negative: a majority indicated that the PC was very or extremely biased, slightly or not at all useful, and slightly or not at all predictive of future performance. The responses to the other critique items reflected combinations of positive, negative, and neutral attitudes, without any attitude representing a majority opinion. However, a plurality of respondents indicated that the PC was very or extremely unfair, aversive, and difficult to complete. Only 31% found the military quality definitions to be either marginally or not at all adequate. Finally, 72% of the respondents were either very or extremely unfavorable toward the implementation of the program.

Discussion

The results of the preliminary administration indicate that the PC technique is a potentially useful procedure for identifying the class members with the highest potential as Army aviation officers, although a majority of the students

were critical of its use. There appeared to be a consensus among the section members about the peers with the highest potential, and the peer assessments did not change over a period of 1 month.

There were, however, several problems with the preliminary administration of the PC procedure. First, the students were not advised about the PC appraisals prior to the first data collection. Several of the class members complained that they would have "acted differently" if they had known of the PC appraisals in advance. Second, a concurrent but surreptitious attempt by the class leaders to evaluate the section members was exposed just before the second data collection. Many of the students were angry that they were being evaluated without their knowledge. Both of these problems may have affected the students' attitudes about the peer evaluations.

Third, many of the students were outprocessing and did not participate in the final data collection and PC critique, especially in Section 2. Fourth, the time that elapsed between the first and second PC data collections was relatively short for evaluating the stability of the peer assessments. Finally, the military qualities and the nominee pairs were presented in a fixed order, which may have resulted in sequence effects. That is, comparisons on subsequent military qualities may have been influenced by previous comparisons, and the respondents may have tended to select the first member in each pair. The first member was always the higher ranked nominee.

Although the results of the preliminary administration were encouraging, the confounding problems were considered serious enough to warrant a more controlled replication. Therefore, a second administration was conducted under controlled experimental conditions to verify the initial results.

SECOND ADMINISTRATION

The second test of the PC procedure was designed as a replication of the preliminary administration, with the following changes:

- students were advised in advance of the research,
- other non-academic student evaluations were prohibited,
- 3 months elapsed between the initial and final data collection, and
- the order of presentation of the military qualities and nominee pairs was counterbalanced.

Method

The general procedures used in the preliminary administration were repeated in the second administration except as noted above. The first PC data sets were collected from AVNOAC class 86-1 during regular class periods at the end of the second month of training. Following an introduction by the Assistant School Secretary, the PC program was described, student questions were answered, and the appraisal materials were distributed. All PC evaluations were completed within 10 minutes.

The second sets of PC data were collected approximately 3 months later, during regular class periods in the last week of the course. The students completed the PC evaluations and critique within 20 minutes. After the graduation of AVNOAC 86-1, FARs were completed by a majority of the faculty advisors and the AVGs were collected from the School Secretary's office.

Results

Usable PC evaluations were collected from 48 students in each section during the first data collection. Incomplete or otherwise unusable PC evaluations were returned by one student in Section 1 and by three students in Section 2. Of the 50 students in Section 1, 36 received nominations by more than one peer; of the 53 students in Section 2, 37 received nominations by more than one peer.

During the second data collection, 47 students in Section 1 and 44 students in Section 2 completed usable PC ratings and student critiques. Five students in Section 2 submitted incomplete or unusable PC evaluations. There were 37 students in Section 1 and 28 students in Section 2 who received nominations from more than one peer.

FAR evaluations were collected from 13 faculty advisors for 89 students (48 in Section 1 and 41 in Section 2). The results were similar to the evaluations for class 85-2: the median percentile was 75 in both classes, with a range of 5-95 in Section 1 and 50-99 in Section 2. The percentile ranks in Section 1 were highly skewed: one student was assigned a rank of 5, three students were assigned a rank of 25, and the rest of the students were assigned ranks of 50 or more. The AVGs were similar in each section of class 86-1 and to the AVGs in class 85-2. The mean AVG was 93.5 (SD = 2.81) in Section 1 and 93.2 (SD = 3.08) in Section 2.

Reliability estimates. Two types of reliability coefficients were computed for the AVNOAC 86-1 PC data. First, the correlations between the initial and final PC scores were .79 in Section 1 (n = 50) and .93 in Section 2 (n = 53), indicating the stability of the PC appraisals across a period of approximately 3 months. The 3-month correlations for the rank scores and each of the military quality comparisons ranged from .72 to .80 in Section 1 and from .91 to .94 in Section 2.

Second, split-half (odd-even) correlations for each data collection in each section were computed to evaluate the internal consistency of the ratings. The correlations were .83 and .86 in Section 1 for the first and second data collections, respectively. The correlations were .96 for both data collections in Section 2. Both sets of correlations are corrected for the reduced number of raters using the Spearman-Brown formula (cf. Downey & Duffy, 1978).

The reliability coefficients are acceptable in all cases, although they are substantially higher in Section 2. Because of the high estimated reliability, the ratings from the two data collections were combined into a single PC score for each peer in each section.

Peer consensus. The combined PC scores ranged from .00 to .24 in Section 1 and from .00 to .47 in Section 2. Two peers in each section received PC scores greater than .20 (see Table 3). As in class 85-2, a majority of the PC scores in both sections were between .00 and .05. The scores indicate a high consensus among the members of Section 2 in identifying the two peers (PC+ = .42 and .47) with the highest potential as aviation officers. The PC scores in Section 1 also identified the peers having the highest potential, although the PC scores were much lower (PC+ = .22 and .24).

Table 3

Frequency Distribution of Peer Comparison (PC) Scores in Sections 1 and 2 of AVNOAC 86-1

PC Range	Section 1			Section 2		
	PC1	PC2	PC+	PC1	PC2	PC+
.000 - .049	31	35	33	40	42	41
.050 - .099	9	5	8	7	5	5
.100 - .149	8	5	5	1	1	2
.150 - .199	0	2	2	3	2	3
.200 - .249	0	3	2	0	1	0
.250 - .299	2	0	0	0	0	0
.300 - .349	0	0	0	0	0	0
.350 - .399	0	0	0	2	0	0
.400 - .449	0	0	0	0	0	1
.450 - .499	0	0	0	0	1	1
.500 - .549	0	0	0	0	0	0
.550 - .599	0	0	0	0	1	0

Note. There were 50 students in Section 1 and 53 students in Section 2. PC1 = first data collection; PC2 = second data collection; PC+ = combined PC scores.

The lower scores in Section 1 could be an artifact of the methodology if there are more than five peers with relatively high potential who are not substantially different from each other. Table 3 shows there were nine students in Section 1 with combined PCs between .10 and .25; in Section 2, there were only five students in this range. That is, when an approximately equal number of points assigned in each section is divided among a larger number of peers, the average PC score will be somewhat lower than if the points were divided among fewer peers.

External correlations. The combined PC scores were then correlated with the FARs and AVGs. For Sections 1 and 2, respectively, the PC correlations were .02 and .30 with the FAR, and .24 and .27 with the AVG. These correlations indicate that the PC score is measuring a different aspect of the class members' performance during the AVNOAC than the FAR and AVG. The .02 correlation between the FAR and PC in Section 1 is partially attributable to the highly skewed distribution of FARs in that section. The correlations

between the FAR and AVG were .53 in both sections. The FAR-AVG correlations probably indicate that the faculty advisors used the academic average as a primary source of rating information.

Internal correlations. The number of nominations, the rank scores, and the number of favorable comparisons on each military quality were correlated to determine whether the components of the combined PC score provided unique information about the section members. As can be seen in Tables 4 and 5, all the component variables are highly correlated. In fact, the PC scores in each section can be perfectly predicted (i.e., $R = 1.0$) without including the number of nominations or the rank score in the regression equation. The multiple regression coefficient is artificially high because each comparison is a component of the PC score. However, the important point is that the regression weights for the military quality values are not uniform: leadership has the highest weight and communication has the lowest weight in both sections (see Table 6).

Table 4

Intercorrelation Matrix of Nomination, Ranking, and Military Quality Comparisons in Section 1

	NOM	RANK	LDRS	RESP	COMM	APPR
RANK	.984					
LDRS	.966	.987				
RESP	.961	.982	.963			
COMM	.957	.954	.959	.924		
APPR	.951	.951	.921	.935	.887	
COOP	.944	.960	.945	.958	.933	.891

Note. NOM = number of nominations; LDRS = leadership; RESP = responsibility; COMM = communication; APPR = appearance; COOP = cooperation.

Table 5

Intercorrelation Matrix of Nomination, Ranking, and Military Quality Comparisons in Section 2

	NOM	RANK	LDRS	RESP	COMM	APPR
RANK	.981					
LDRS	.973	.991				
RESP	.960	.978	.965			
COMM	.955	.981	.978	.981		
APPR	.937	.970	.954	.929	.945	
COOP	.965	.967	.957	.986	.965	.895

Note. NOM = number of nominations; LDRS = leadership; RESP = responsibility; COMM = communication; APPR = appearance; COOP = cooperation.

Table 6

Regression weights to Predict Peer Comparison Scores in Sections 1 and 2 of AVNOAC 86-1

Variable	Section 1	Section 2
Leadership	1.62	1.55
Responsibility	1.37	1.14
Communication	1.07	.93
Appearance	1.25	1.44
Cooperation	1.14	1.39
Constant	1.09	1.10

Note. $R = 1.0$ in each section, $p < .0001$.

PC critique. Finally, the PC critique responses from class 86-1 were negative overall, but not as negative as those from class 85-2. The reactions to the PC program were generally very similar in each section of class 86-1 (see Appendix B). A majority of the respondents indicated that the PC was either slightly or not at all useful for selecting AVNOAC honor graduates. There were greater differences in the opinions held by the respondents on the issues of PC

fairness, bias, and predictability of future Army performance. Ratings of the adequacy of definitions and the difficulty of nominating, ranking, and comparing peers were very similar to the 85-2 results. Despite the slight positive shift in attitude toward the PC program, approximately 69% of the respondents were still either very or extremely unfavorable toward the implementation of the PC procedure.

Discussion

The results of the 86-1 PC administration support the conclusion drawn from the 85-2 results: the PC procedure is a potentially useful method for identifying the peers with the highest potential as Army aviation officers, at least in terms of the reliability of the ratings. There was a consensus about the peers with the highest potential and the ratings were generally stable over a 3-month data collection interval. However, longitudinal research is required to determine the validity of the PC technique for predicting future performance.

The high correlations between the nomination, ranking, and comparison variables may indicate a halo effect. That is, the initial choices made by each section member may have influenced his or her subsequent judgments. This interpretation would indicate that the members were not effectively discriminating differences between their peers on the various components of the PC score. However, if the military quality dimensions are the primary components of the overall nomination and ranking process, then the high intercorrelations represent a high level of internal consistency for the total scale. That is, each of the military quality comparisons are homogeneous "items" that increase the reliability of the total scale. The differential weights for the five military quality comparisons in the multiple regression analyses support an interpretation that the section members were reliably discriminating among their peers on highly correlated variables.

Whether or not the military quality comparisons are redundant and superfluous because of halo effect or are important contributors to the overall reliability of the evaluations, the military quality comparisons do serve two valuable functions. First, the military qualities provide a common frame of reference about the characteristics that senior aviation officers believe are the most critical for current and advanced levels of performance. Second, the additional complexity of the procedure (compared to simple

nomination or ranking procedures) makes it more difficult for disgruntled section members to game or subvert the appraisal process.

Similar to the 85-2 students, the 86-1 students found the rating procedure to be aversive and were unfavorable toward the implementation of the PC technique. This finding is consistent with the majority of other research results that have evaluated user reaction to peer appraisals (e.g., Love, 1981). However, the same results have usually demonstrated that the peer appraisals are highly reliable and valid. Advance notification of the PC evaluation and an emphasis on the positive rather than negative nomination aspect (i.e., identifying excellent rather than unacceptable performers) may mitigate user reaction if it is established as part of the AVNOAC evaluation procedures.

UTILIZATION OPTIONS

When this research was initiated, the student in each class with the highest academic average was named the distinguished graduate and the four students with the next highest averages were named honor graduates. The top 20% of the class in academic average were named to the Commandant's list. For the top 20% of AVNOAC 86-1 (n = 21), the academic averages are near 100, with a range of only 2.67 points (see Table 7). Approximately 83% of the 103 students in the class had academic averages greater than 91; the lowest average was 83.78. That is, the majority of the class exhibited excellent academic performance during the course. Conversely, the academic averages do not discriminate very well among the top performers: the averages for the students ranked ninth and tenth are tied at two decimal places.

If the PC procedure is implemented in the AVNOAC, a method will be required for combining the academic and PC criteria to select students for class honors. There are two primary options for combining the two types of evaluative data: a multiple gate approach and a weighted sum approach.

Multiple Gate Approach

A multiple gate approach would require a member to be in the top percentiles on the whole person criterion (i.e., the PC evaluation) to be eligible for honors. The cutoff percentile could be set to allow either a small or a large percentage of the class members to be eligible for honors. Increasing the percentile cutoff would increase the importance of the PC evaluation in relation to the academic evaluation. Once eligibility was established, honors would be awarded solely on the basis of the academic criterion.

For example, if a PC criterion were set at .05, nine of the top academic students would not be eligible for honors. In their place, students would be selected down to an academic ranking of 48. The last student included on the Commandant's list would have both a high AVG (93.99) and a moderately high PC score (.12). The last student included on the Commandant's list under only the academic criterion would have a slightly higher AVG (96.11) but a much lower PC score (.01).

Table 7

Alternative Combinations of Academic and Peer Comparison Scores in AVNOAC 86-1

SEC	AVG	PC	RANK	RANK1	RANK5	RANK10
2	98.78	.47	1	1	1	1
1	98.35	.18	2	2	2	2
1	98.11	.04	3	3	4	6
2	97.67	.16	4	4	3	3
1	97.62	.12	5	5	5	4
1	97.56	.04	6	6	7	9
2	97.27	.05	7	8	9	11
2	97.22	.14	8	7	6	5
1	97.10	.03	9	10	12	13
1	97.10	.05	10	9	11	12
1	96.94	.10	11	11	10	8
1	96.90	.04	12	12	13	15
2	96.75	.15	13	13	8	7
2	96.57	.02	14	14	16	19
1	96.36	.04	15	16	18	20
1	96.33	.10	16	15	15	14
2	96.29	.03	17	18	20	21
1	96.24	.01	18	19	21	23
2	96.22	.17	19	17	14	10
1	96.13	.07	20	20	19	18
1	96.11	.01	21	22	23	24
2	96.04	.11	24	21	17	16
1	94.43	.24	41	38	28	17

Note. AVNOAC = Aviation Officer Advanced Course; SEC = section; AVG = academic average; PC = peer comparison score; RANK = AVG rank order; RANK1 = rank order of AVG + 1(PC); RANK5 = rank order of AVG + 5(PC); RANK10 = rank order of AVG + 10(PC). The dashed lines indicate a break in the rank ordering of the AVGs.

Weighted Sum Approach

The weighted sum approach combines the PC score and the academic average using predetermined weights to produce a single criterion for awarding class honors. Both the AVG and PC raw scores are proportions (i.e., percentage of maximum

possible scores) but the AVG is usually expressed with a weight of 100. That is, the AVG can range from 0 to 100 with fractional values. The effect of using different weights for the PC scores is demonstrated in Table 7.

In Table 7, the RANK column presents the rank order of the top students using only the academic criterion. RANK1 is the rank order of the students using a weight of 100 for AVG and 1 for PC. RANK5 is the rank order of the students using a weight of 100 for AVG and 5 for PC. RANK10 is the rank order of the students using a weight of 100 for AVG and 10 for PC.

With a PC weight of 1, the rank order of the top students changes very little from the academics-only rank order. There are minor shifts of one or two ranks among the students having academic ranks of 7 through 19. The only change that would affect an honors award involves the last student in the top 20% academically. The last student would be dropped from the Commandant's list and would be replaced by the student with an academics-only rank of 24.

With a PC weight of 5, there is a maximum change of five places in rank order but no further changes in the students receiving awards. With a PC weight of 10, however, there is one change in the students named as honor graduates and a second student is dropped from the Commandant's list. The latter student is replaced by a student with an academic rank of 41 who received the highest PC score in Section 1 (.24). The high PC score with a weight of 10 resulted in a change of 24 places in rank order.

Table 7 demonstrates the specific results that would occur under various options in AVNOAC 86-1. Many other options are possible and the same options may produce different results in other classes. For example, if the PC score were given a weight of 39 or higher, the student with the second highest PC score (.42) would change from an academic rank of 90 to a combined rank of 2. In class 86-1, the student with the highest PC (.47) also had the highest academic average. No combination of AVG and PC scores would result in a change in the distinguished graduate. However, if that student had had a slightly lower AVG of 97.00, the PC score would change his rank from tenth to first with a PC weight of 5.

SUMMARY AND RECOMMENDATIONS

The results of this research indicate that the PC procedure is a highly reliable and easy-to-use method for evaluating the AVNOAC students with the highest potential as Army aviation officers. The PC procedure exhibited high levels of internal consistency reliability and temporal stability over periods of 1 and 3 months. In both administrations, there was a clear consensus among the section members about the peers with the highest potential. Prior research (e.g., Downey, Medland, & Yates, 1976; Eastman & Leger, 1978; Gilbert & Downey, 1978; Kane & Lawler, 1978) suggests that the results are likely to be valid as well, but longitudinal research is required to determine the predictive validity of the PC procedure.

In addition to identifying students with high career potential, the whole-person honors criterion was intended to motivate the students to exercise their military qualities as well as their academic abilities during the AVNOAC. Although data were not collected to evaluate this objective directly, anecdotal reports from students in classes 85-2 and 86-1 indicated that the peer evaluations caused many of the class members to improve their military decorum during the course.

The greatest disadvantage of the procedure is the negative reaction from the majority of the students. Comments from these students most often expressed concerns for the "popularity contest" issue and the potential for gaming the procedure. However, previous research has shown that popularity and friendship do not affect the reliability and validity of peer assessments (e.g., Love, 1981). Although the PC procedure is easy to use, it is sufficiently complex to render it difficult to game. Many of the students also expressed doubts that the evaluations were for research purposes only; these doubts may have influenced their overall reaction to the procedure.

Not all the students were opposed to the procedure and a few strongly favored its implementation. These students typically cited the need for evaluations of other than academic performance and recognized that the AVNOAC class members had the best opportunity to observe and evaluate their peers. However, the appropriateness of the PC procedure should be reviewed if there are major changes in the structure of the AVNOAC.

Implementation

Despite the negative peer reactions and the lack of specific validity information, the research results support the implementation of the PC program in the AVNOAC course. Specifically, it is recommended that the program be implemented on a 1-year trial basis with the PC score assigned a relatively low weight in the selection algorithm. Data should be collected during the trial period to evaluate the effect of the PC score on the selection of students for honors and the reaction of the peers toward the operational use of the PC procedure.

This information should provide a more stable data base for deciding whether to terminate the PC procedure, implement it on a permanent basis, or modify it before implementation. If the trial implementation in the AVNOAC is successful, the need for the PC procedure in other courses, such as the Aviation Officer Basic Course, can be evaluated. The data collected during the 1-year trial can also serve as a baseline for the longitudinal validation of the PC procedure.

In addition, the following technical recommendations are made for implementing the PC procedure during the trial period:

- The students should be advised during orientation that the peer evaluations will be conducted and told how the criteria will be combined.
- The PC evaluations should first be collected after approximately 6 to 8 weeks. The initial administration is designed to familiarize the students with the PC procedure and the evaluation dimensions, and to provide data for evaluating the temporal reliability of the procedure.

- The final PC evaluations and critique should be collected during the last 2 weeks of the course but before the final academic examination is administered.

- If the initial and final data collections are highly correlated, the PC scores should be combined. If they are not highly correlated, only the final PC data should be used.

- For the trial implementation, the weighted sum combination is recommended because of the greater interaction between each student's academic average and PC score. The multiple gate is a viable alternative, but it converts the PC score into a

dichotomous variable. That is, students with very high PC scores are not distinguished from students with PC scores that are slightly above the minimum.

REFERENCES

Bernardin, H. J., & Beatty, R. W. (1984). Performance appraisal: Assessing human behavior at work. Boston, MA: Kent Publishing.

Bolton, W. L. (1971). An application of constant sum paired comparison technique to peer ratings at an Army ROTC summer camp. Unpublished Masters thesis. Knoxville, TN: University of Tennessee, Knoxville.

Burke, W. P., Kramer, R. C., & Butler, R. P. (1982). Development of performance-based assessment centers for admissions officers at the U.S. Military Academy (Research Report 1339). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. **(AD A139 816)**

Cederblom, D., & Lounsbury, J. W. (1980). An investigation of user acceptance of peer evaluations. Personnel Psychology, 33, 567-579.

Cox, J. A., & Krumboltz, J. D. (1958). Racial bias in peer ratings of basic airmen. Sociometry, 21, 292-299.

deJung, J. E., & Kaplan, H. (1962). Some differential effects of race of rater and ratee on early peer ratings of combat aptitude. Journal of Applied Psychology, 46, 370-374.

DeNisi, A. S., & Mitchell, J. L. (1978). An analysis of peer ratings as predictors and criterion measures and a proposed new application. Academy of Management Review, 3, 369-374.

Downey, R. G., Medland, F. F., & Yates, L. G. (1976). Evaluation of a peer rating system for predicting subsequent promotion of senior military officers. Journal of Applied Psychology, 61, 206-209.

Downey, P. G., & Duffy, P. J. (1978). Review of peer evaluation research (Technical Paper 342). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. **(AD A061 780)**

Eastman, R. F., & Leger, M. (1978). Validity of associate ratings of performance potential by Army aviators (Research Memorandum 78-24). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A077 971)

Eastman, R. F., & McMullen, R. L. (1976). Reliability of associate ratings of performance potential by Army aviators (Research Memorandum 76-28). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A076 826)

Engen, T. (1971). Psychophysics II: Scaling methods. In J. W. Kling and L. A. Riggs (Eds.), Experimental psychology (3rd ed.) (pp. 51-54). New York: Holt, Rinehart, and Winston.

Gilbert, A. C. F., & Downey, R. G. (1978). Validity of peer ratings obtained during ranger training (Technical Paper 344). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A061 576)

Griffin, G. R., & Mosko, J. D. (1977). Naval aviation attrition 1950-1976: Implications for the development of future research and evaluation (Report 1237). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer supervisor ratings. Personnel Psychology, 41, 43-62.

Hollander, E. P. (1954). Buddy ratings: Military research and industrial implications. Personnel Psychology, 7, 385-393.

Hollander, E. P. (1956). The friendship factor in peer nominations. Personnel Psychology, 9, 435-447.

Hollander, E. P. (1957). The reliability of peer nominations under various conditions of administration. Journal of Applied Psychology, 41, 85-90.

Hollander, E. P., & Webb, W. B. (1955). Leadership, followership, and friendship: An analysis of peer nominations. Journal of Abnormal and Social Psychology, 50, 163-167.

Kane, J. S., & Lawler, E. E. III. (1978). Methods of peer assessment. Psychological Bulletin, 85, 555-586.

Korman, A. K. (1968). The prediction of managerial performance: A review. Personnel Psychology, 21, 295-322.

Lawshe, C. H., Jr., Kephart, N. C., & McCormick, E. J. (1949). The paired comparison technique for rating performance of industrial employees. Journal of Applied Psychology, 33, 69-77.

Lewin, A. Y., & Zwany, A. (1976). Peer nominations: A model, literature critique and a paradigm for research. Personnel Psychology, 29, 423-447.

Love, K. G. (1981). Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. Journal of Applied Psychology, 66, 451-457.

Mayfield, E. C. (1970). Management selection: Buddy nominations revisited. Personnel Psychology, 23, 377-391.

McAnulty, D. M., & Jones, D. H. (1984). An evaluation of aviator training ability requirements scale ratings. In M. J. Alluisi, S. deGroot, and E. A. Alluisi (Eds.), Proceedings of the 28th Annual Meeting of the Human Factors Society (pp. 356-361). Santa Monica, CA: The Human Factors Society.

McEvoy, G. M., & Buller, P. F. (1987). User acceptance of peer appraisals in an industrial setting. Personnel Psychology, 40, 785-797.

Puryear, E. F. (1971). Nineteen stars: A study in military character and leadership. Washington, DC: Coiner Publications.

Roadman, H. E. (1964). An industrial use of peer ratings. Journal of Applied Psychology, 48, 211-214.

Rogers, R. W., Lilley, L. W., Wellins, R. S., Fischl, M. A., & Burke, W. P. (1982). Development of the precommissioning leadership assessment program (Technical Report 560). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A130 927)

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413-428.

Siegel, L. (1982). Paired comparison evaluations of managerial effectiveness by peers and supervisors. Personnel Psychology, 35, 843-852.

Schmidt, F. L., & Johnson, R. H. (1973). Effect of race on peer ratings in an industrial situation. Journal of Applied Psychology, 57, 237-241.

Wahlberg, J. L., Boyles, W. R., & Boyd, H. A. (1971). Peer ratings as predictors of success in military aviation (Professional Paper 1-71). Alexandria, VA: Human Resources Research Organization.

Waters, L. K., & Waters, C. W. (1970). Peer nominations as predictors of short-term sales performance. Journal of Applied Psychology, 54, 42-44.

A P P E N D I X A

THE AVNOAC PEER COMPARISON CRITIQUE FORM

Please complete the critique on the use of the peer comparison technique in the AVNOAC. Read each question and the response alternatives, then check the response that most accurately reflects your views. The final item provides space for additional comments, criticisms, or recommendations. The critique should be submitted anonymously.

1. How FAIR is the peer comparison technique for use in the AVNOAC? That is, do all class members have an equal opportunity to be identified as outstanding by their peers?

[ ]	[ ]	[ ]	[ ]	[ ]
EXTREMELY FAIR	VERY FAIR	SOMEWHAT FAIR	VERY UNFAIR	EXTREMELY UNFAIR

2. How BIASED do you think the peer comparisons will be? That is, do you think that friendship, race, sex, or other factors will influence the comparisons?

[ ]	[ ]	[ ]	[ ]	[ ]
EXTREMELY BIASED	VERY BIASED	SOMEWHAT BIASED	SLIGHTLY BIASED	NOT AT ALL BIASED

3. How USEFUL do you think the peer comparisons will be in selecting the class honor graduates? That is, do you think the results will be valid indicators of the most outstanding class members?

[ ]	[ ]	[ ]	[ ]	[ ]
EXTREMELY USEFUL	VERY USEFUL	SOMEWHAT USEFUL	SLIGHTLY USEFUL	NOT AT ALL USEFUL

4. How PREDICTIVE do you think the peer comparisons will be in identifying class members with the highest potential as U.S. Army aviation officers? That is, do you think the results will be valid indicators of the individuals who are most likely to have highly successful Army careers?

[ ]	[ ]	[ ]	[ ]	[ ]
NOT AT ALL PREDICTIVE	SLIGHTLY PREDICTIVE	SOMEWHAT PREDICTIVE	VERY PREDICTIVE	EXTREMELY PREDICTIVE

5. How AVERSIVE is the peer comparison procedure to you? That is, how much do you resent having to make the peer comparisons?

[ ]	[ ]	[ ]	[ ]	[ ]
NOT AT ALL AVERSIVE	SLIGHTLY AVERSIVE	SOMEWHAT AVERSIVE	VERY AVERSIVE	EXTREMELY AVERSIVE

6a. How ADEQUATE are the military quality definitions? That is, do the military quality definitions provide a common frame of reference for making the peer comparisons?

[ ]	[ ]	[ ]	[ ]	[ ]
NOT AT ALL ADEQUATE	MARGINALLY ADEQUATE	FAIRLY ADEQUATE	VERY ADEQUATE	EXTREMELY ADEQUATE

6b. Check the military qualities, if any, that have inadequate definitions.

[ ]	[ ]	[ ]	[ ]	[ ]
LEADERSHIP	RESPONSI-BILITY	COMMUNICA-TION	APPEARANCE	COOPERATION

7. How difficult was it to identify the five members of your section with the highest potential as U.S. Army aviation officers?

[ ]	[ ]	[ ]	[ ]	[ ]
NOT AT ALL DIFFICULT	SLIGHTLY DIFFICULT	SOMEWHAT DIFFICULT	VERY DIFFICULT	EXTREMELY DIFFICULT

8. How difficult was it to rank order the five members you identified?

[ ]	[ ]	[ ]	[ ]	[ ]
NOT AT ALL DIFFICULT	SLIGHTLY DIFFICULT	SOMEWHAT DIFFICULT	VERY DIFFICULT	EXTREMELY DIFFICULT

9a. How difficult was it to compare the peers on the military qualities?

[ ]	[ ]	[ ]	[ ]	[ ]
NOT AT ALL DIFFICULT	SLIGHTLY DIFFICULT	SOMEWHAT DIFFICULT	VERY DIFFICULT	EXTREMELY DIFFICULT

9b. Check the military qualities, if any, that were extremely difficult to judge.

[ ]	[ ]	[ ]	[ ]	[ ]
LEADERSHIP	RESPONSI-BILITY	COMMUNICA-TION	APPEARANCE	COOPERATION

10. How FAVORABLE are you toward using the peer comparisons in the AVNOAC? That is, do you think the peer comparisons should be used in the AVNOAC or not?

[ ]	[ ]	[ ]	[ ]	[ ]
EXTREMELY UNFAVORABLE	SOMEWHAT UNFAVORABLE	INDIFFERENT	SOMEWHAT FAVORABLE	EXTREMELY FAVORABLE

11. When should the peer comparisons be collected during the AVNOAC?

[ ]	[ ]	[ ]	[ ]
ONE MONTH AFTER THE COURSE BEGINS	AT THE MID-POINT OF THE COURSE	ONE MONTH BEFORE GRADUATION	ONE WEEK BEFORE GRADUATION

12. Additional comments, criticisms, or recommendations:_____

Thank you for your assistance. The critique should be submitted anonymously.

A P P E N D I X B

PEER COMPARISON CRITIQUE RESULTS FOR AVNOAC CLASS 86-1

Critique Dimension	Section 1	Section 2
1. Fairness		
- % extremely unfair	15.2	12.3
- % very unfair	30.4	28.6
- % somewhat fair	37.0	40.8
- % very fair	13.0	14.3
- % extremely fair	4.4	4.1
2. Bias		
- % extremely biased	26.1	20.4
- % very biased	15.2	32.7
- % somewhat biased	34.8	34.7
- % slightly biased	15.2	8.2
- % not at all biased	8.7	4.1
3. Useful		
- % not at all useful	45.7	28.6
- % slightly useful	23.9	28.6
- % somewhat useful	21.7	32.7
- % very useful	8.7	8.2
- % extremely useful	0.0	2.0
4. Predictive		
- % not at all predictive	34.8	12.3
- % slightly predictive	23.9	26.5
- % somewhat predictive	30.4	44.9
- % very predictive	8.7	14.3
- % extremely predictive	2.2	2.0

5. Aversive		
- % extremely aversive	26.1	26.5
- % very aversive	15.2	8.2
- % somewhat aversive	26.1	26.5
- % slightly aversive	17.4	16.3
- % not at all aversive	15.2	22.5
6a. Adequacy of definitions		
- % not at all adequate	8.7	8.2
- % marginally adequate	19.6	28.6
- % fairly adequate	47.8	32.7
- % very adequate	19.6	26.5
- % extremely adequate	4.4	2.0
6b. Percent listing each definition as inadequate[a]		
- Leadership	34.8	38.8
- Responsibility	28.3	12.3
- Communication	15.2	10.2
- Appearance	10.9	14.3
- Cooperation	23.9	10.2
- None indicated	52.2	40.8
7. Difficulty to nominate five peers		
- % extremely difficult	21.7	10.2
- % very difficult	28.3	26.5
- % somewhat difficult	21.7	24.5
- % slightly difficult	6.5	18.4
- % not at all difficult	21.7	18.4
- % no response	0.0	2.0
8. Difficulty to rank order five peers		
- % extremely difficult	23.9	16.3
- % very difficult	34.8	36.7
- % somewhat difficult	19.6	12.3
- % slightly difficult	10.9	26.5
- % not at all difficult	10.9	8.2

9a. Difficulty to compare five peers		
- % extremely difficult	13.0	16.3
- % very difficult	23.9	34.7
- % somewhat difficult	30.4	26.5
- % slightly difficult	19.6	14.3
- % not at all difficult	13.0	8.2
9b. Percent extremely difficult to compare[a]		
- Leadership	63.0	67.4
- Responsibility	50.0	65.3
- Communication	17.4	10.2
- Appearance	15.2	10.2
- Cooperation	32.6	28.6
- None indicated	19.6	8.2
10. Favorable to implementation		
- % extremely unfavorable	50.0	42.9
- % very unfavorable	19.6	26.5
- % indifferent	17.4	14.3
- % very favorable	4.4	10.2
- % extremely favorable	6.5	6.1
- % no response	2.2	0.0
11. When to administer		
- % 1 month after beginning	0.0	4.1
- % at course midpoint	4.4	4.1
- % 1 month before end	23.9	28.6
- % 1 week before end	37.0	34.7
- % other (e.g., never, twice)	17.4	20.4
- % no response	17.4	8.2

Note. There were 46 critiques received in Section 1 and 49 critiques received in Section 2. All critiques were submitted anonymously. Some items may not add to 100% due to rounding error.

[a]Multiple responses were permitted to these items; the totals for each section may add to more than 100%.